View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Universiteit Twente Repository

Internet Engineering Task Force
INTERNET-DRAFT
Expires April 2003

L. Westberg
G. Heijenk
G. Karagiannis
S. Oosthoek
D. Partain
V. Rexhepi
R. Szabo
P. Wallentin
Ericsson
Hamad el Allali
University of Twente
October 2002

Resource Management in Diffserv Measurement-based Admission Control PHR

draft-westberg-rmd-mbac-phr-00.txt

Status of this Memo

This document is an Internet-Draft and is in full conformance with all provisions of Section 10 of RFC2026.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts. Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

Distribution of this memo is unlimited.

Copyright Notice

Copyright (C) The Internet Society (2001). All Rights Reserved.

1. Abstract

The purpose of this draft is to present the Resource Management in Diffserv (RMD) Measurement-Based Admission Control (RIMA) Per Hop Reservation (PHR) protocol. The RIMA PHR protocol is used on a per-hop basis in a Differentiated Services (Diffserv) domain and extends the Diffserv Per Hop Behavior (PHB) with Measurement-based Admission Control features.

Table of Contents

2. Introduction

The current definition of Diffserv [RFC2475] does not contain a simple and scalable solution to the problem of measurement-based admission control. The Resource Management in Diffserv (RMD) measurement-based admission control (RIMA) Per Hop Reservation (PHR) protocol presented in this document operates in an edge-to-edge Diffserv domain extending the Per Hop Behavior (PHB) functionality with measurement-based admission control (MBAC) features.

The RIMA PHR is a unicast edge-to-edge protocol that is applied in a Diffserv domain and aims at extreme simplicity and low cost of implementation along with good scaling properties. The RIMA PHR protocol operates on a hop-by-hop basis on all nodes, both edge and interior, located in an edge-to-edge Diffserv domain. This PHR protocol can be applied in Diffserv domains that are using either of the two Internet Protocol (IP) versions (version 4 [RFC791] or version 6 [RFC2460]).

The edge and interior nodes used in the RIMA PHR do not maintain any aggregated reservation state. However, each node MUST be able to observe the traffic load status of each DSCP class by measuring the traffic (user) data load per DSCP class.

The Resource Management in Diffserv (RMD) Framework document [RMD-frame] specifies how a MBAC PHR can interoperate with a Per Domain Reservation (PDR) protocol. A PDR scheme represents the resource reservation in the Diffserv domain, and it is implemented only at the boundary of the domain (at the edge nodes).

3. Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Furthermore, all the new terms used in this draft are to be interpreted as described in [RMD-frame].

4. RIMA PHR functionality

The RIMA PHR protocol performs the following functions.

* Stores a pre-configured threshold value on maximal allowable resource units per PHB.

* Determination of traffic load status. Each node MUST be able to determine the traffic load status of each DSCP class by measuring the traffic (user) data load per DSCP class. The "traffic load status" specifies how many resource units allocated to a particular DSCP class are in use.

* Detection and notification of severe congestion. Severe congestion can be considered as an undesirable state which may occur as a result of a route change or a link failure. Typically, routing algorithms are able to adapt and change their routing decisions to reflect changes in the topology and traffic volume. In such situations the re-routed traffic will have to follow a new path. Nodes located on this new path may become overloaded, since they suddenly might need to support more traffic than their capacity. All nodes MUST be able to identify a severe congestion situation. The RIMA PHR protocol provides the means of informing other nodes of the congestion situation on a hop-by-hop basis.

* Adaptation to load sharing. Load sharing allows interior nodes to take advantage of multiple routes to the same destination by sending via some or all of these available routes. The RIMA PHR protocol has to adapt to load sharing once it is used.

* Transport of transparent PDR messages. The PHR protocol may encapsulate and transport PDR messages sent from an ingress node to an egress node.

5. RIMA PHR protocol operation

There are two main RIMA PHR protocol operations:

* normal operation, which refers to the situation when no performance degradation problems are occurring in the network.

* fault handling, which refers to the situations when there are performance degradation problems in the network, such as route or link failures. These situations may result in

severe congestion occurrence or loss of PHR signaling messages.

5.1. RIMA PHR Protocol Messages

In RIMA, only one PHR protocol messages is specified: the "PHR_Resource_Request". This PHR message will pass through the same nodes as the actual data traffic will pass through.

5.1.1. PHR_Resource_Request

The "PHR_Resource_Request" is used to determine the resource utilization status for each PHB, on all nodes located on the communication path between the ingress and egress nodes according to an external QoS Request. This status represents the current traffic load status for each PHB that is determined by means of measurements of the (average) rate of the traffic load. The ingress node generates for each new incoming flow a "PHR_Resource_Request" message, which signals only the resource units requested by this particular flow. The acceptance/rejection of this resource request will be decided by an Measurement Based Admission Control Algorithm (MBAC) algorithm.

5.2. RIMA PHR Normal operation

A single RIMA PHR protocol message is specified: the "PHR_Resource_Request". This message passes through the same nodes as the actual traffic will pass through.

The "PHR_Resource_Request" PHR protocol message is sent by an ingress node towards an egress edge node and is used for requesting resources at each node located in the communication path between the ingress and egress nodes.

Any node that receives a PHR protocol message ("PHR_Resource_Request") MUST identify the DSCP type of these signaling packets. Subsequently, a Measurement Based Admission Control Algorithm (MBAC) has to be used in order to admit or reject the request.

An example of a MBAC is the following. If the sum of the value of the PHR Requested Resources (RR) and the value specified by the traffic load (TL)status is less than or equal to the maximum node

capacity or threshold (TH) associated with the given DSCP, i.e., RR + TL <= TH then the request is accepted. Otherwise, i.e., RR + TL > TH, the request is rejected and this packet is marked by setting the "M" bit to the value of "1".

5.3. Fault handling operation

When a node detects severe congestion, it MUST inform the egress node by setting the "S" field of any received PHR message to "1" and sending this message towards the egress node. If this is not possible, operational management solutions, such as Simple Network Management Protocol (SNMP) notifications SHOULD be used to signal severe congestion to the edges.

Moreover, when an interior node detects this situation, it SHOULD notify the egress node by using DSCP remarking of user data packets that are passing through the node. Proportionally to the detected overload, the interior node will remark a number of user data packets which are passing through a severe congested interior node and are associated to a certain PHB, into a domain specific DSCP (see [RFC2474]). [RMD-frame] describes a severe congestion handling procedure which uses the DSCP remarked packets and solves the severe congestion situation.

Any "S" marked (the "S" bit is 1) "PHR_Resource_Request" messages that arrives in an interior node are not processed and are forwarded untouched.

6. RIMA PHR message formats

The PHR protocol information is carried in:

* an IP header Options field, as defined in the [RFC791], when IPv4 is used

* an option field encoded into the Hop-by-Hop Options Extended Header, as defined in [RFC2460], when IPv6 is used

We denote this IP Option field as the RIMA PHR option.

6.1. Message Format in IPv4

The PHR protocol messages used in IPv4 Diffserv domains are represented by the combination of the DSCP field and the contents of an IPv4 option header field [RFC791]. This IPv4 option header field has the following format.

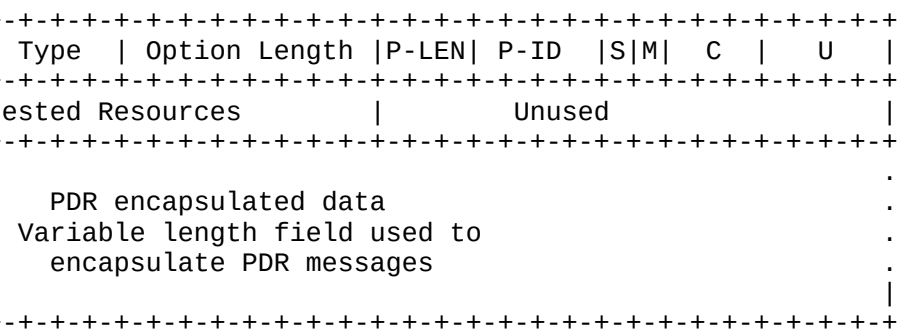

Figure 1: PHR Option field in the IPv4 Option header field

Option Type	8-bit identifier of the type of option. The semantics of this field are specified in [RFC791].
Option Length	8-bit field. This is specified in [RFC791] and represents the length of the Option-Data field of this option, in octets. The option data field consists of all fields included in the option field of the IPv4 header and are placed after the "Option Length" field.
P-LEN (PHR length)	3-bit field. This specifies the length in octets of the specific PHR information data included in the "Option-Data" field. This information does not include the encapsulated PDR information. The value 0 specifies that this IP option field contains only PDR data and no PHR data. The PDR data MUST begin on the next 32-bit word boundary (after the first "unused" field). In this case, the sender MUST set the

	"S", "M", "C", and "unused" fields to 0. The P-ID MUST have the value 2. If a node receives a packet with a P-LEN value of 0, it MUST ignore the values in the "S", "M", "C", and "unused" fields.
P-ID (PHR type)	4-bit field. This specifies the PHR type. For this memo, the value MUST be 2 (RIMA PHR).
S (Severe Congestion)	1-bit field. This field is set to 1 by an interior or edge node when a severe congestion situation is detected. Otherwise, this value is set to 0.
M (Marked)	1-bit field. This field is set to 1 by an interior or edge node when the node cannot satisfy the "Requested Resources" value. Otherwise this value is set to 0.
C (Message type)	3-bit field. This field specifies the type of the PHR message. C Description --- 0 Reserved 1 "PHR_Resource_Request" 2-7 Unused
U(Unused	4-bit currently unused field. Reserved for future PHR extensions.
Requested Resources	16-bit field. This field specifies the requested number of resource units to be reserved by a node. The unit is not necessarily a simple bandwidth value. It may be defined in terms of any resource unit (e.g., effective bandwidth) to support statistical multiplexing at message level.

Unused	16-bit currently unused field. Reserved for future PHR extensions.
PDR encapsulated data	PDR encapsulated information data. This field is only processed by the edge nodes.

6.2. Message Format in IPv6

The PHR protocol messages used in IPv6 Diffserv domains are represented by the combination of the DSCP field and the contents of an option field of a IPv6 Hop-by-Hop header option [RFC2460]. This IPv6 option field has the following format.

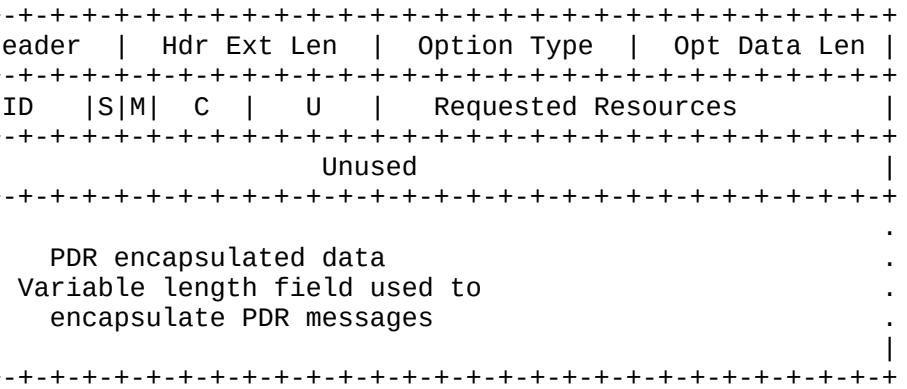

Figure 2: PHR Option field in the IPv6 Hop-by-Hop Header Option

Next Header	8-bit selector. This is specified in [RFC2460] and identifies the type of header immediately following the Hop-by-Hop Options header.
Hdr Ext Len	8-bit field. This is specified in [RFC2460] and represents the length of the Hop-by-Hop Options header in 8-octet units, not including the first 8 octets.
Option Type	8-bit identifier of the type of option. The semantics of this field are specified in [RFC2460].

Opt Data Len	8-bit field. This is specified in [RFC2460] and represents the length in octets of the Option Data field of this option. The option data field consists of all fields included in the Hop-by-Hop header option and placed after the "Opt Data Len" field.
P-LEN (PHR length)	3-bit field. The semantics of this field are identical to the field in the IPv4 option. Like in IPv4, the value 0 specifies that this IP option field contains only PDR data and no PHR data. The PDR data MUST begin on the next 32-bit word boundary (after the first "Requested Resources" field). In this case, the sender MUST set the "S", "M", "C", "unused", and "Requested Resources" fields to 0. The P-ID MUST have the value 1. If a node receives a packet with a P-LEN value of 0, it MUST ignore the values in the "S", "M", "C", "U", and "Requested Resources" fields.
U	4-bit currently unused field. Reserved for future PHR extensions.
Unused	32-bit field that is currently unused. Reserved for future PHR extensions.
PDR encapsulated data	a variable length field that contain PDR encapsulated information data. This field is only processed by the edge nodes.

The "Requested Resources", "P-ID", "S", "M" and "C" fields in Figure 2 are identical to those shown in Figure 1.

7. Adaptation for load sharing

This section is identical to Section 6 presented in [RODA]. Due to load sharing (see e.g., [RFC2676]) a route cycles between different routes in order to balance the load. This will imply that the traffic (user) data may not follow exactly the same paths as the PHR messages used to reserve the transport resources used by this traffic (user) data. As such, interior and edge nodes MUST be able to observe when a load sharing situation occurs.

It is recommended that interior and edge nodes SHOULD forward the PHR messages in such a way that they will follow the same forwarding path as the traffic (user) data associated with these PHR messages. When this cannot be done, we propose use of the same solutions as the multi-path route solutions proposed in Section 1.4.6 of [BaIt00].

These are:

* the data may be tunneled from the ingress to egress node using technologies such as IP-in-IP, GRE (Generic Routing Encapsulation), MPLS (Multiple Label Protocol Switching) label-switched paths, and so on.

* measurement could be used to determine what proportion of traffic for a given reservation travels along each of the load sharing paths, thereby verifying that there is sufficient bandwidth for the reservation.

* by reserving the total capacity of the route down each load sharing path.

In case a network domain is using a routing protocol which is applying an equal cost load sharing principle, any interior node SHOULD be able to know the number, e.g., "N", of multiple equal cost paths that the routing protocol will use to provide the load sharing principle. Subsequently, for each arrived PHR message which is affected by the load sharing principle, the interior node SHOULD be able to create "N" number of PHR messages of identical type as the original one. Each of these generated PHR messages SHOULD contain in its "Requested Resources" field a value equal to the requested resources value which was included in the "Requested Resources" field of the original PHR message divided by the number of equal cost paths, i.e., "N". Moreover, each of these generated PHR messages SHOULD also contain in its "Shared %" field a new value that is calculated by dividing the shared percentage value, included in the

"Shared %" field of the original PHR message, by the number of equal cost paths, i.e., "N".

8. Accuracy of measurements

Since the RIMA PHR is measurement-based, accurate measurements of the available resources during certain time periods are necessary for achieving high utilization of the network. We denote these measurement time periods as measurement periods.

In [BrJa00] several simulation results emphasize that the method of measuring the traffic load status is not significant. This can be proven by the fact that different algorithms that are used to measure the traffic load status can achieve an identical level of performance. The right tuning of knobs of the measurement mechanism, e.g., measuring window size W, sampling time S, is needed to achieve good measurement accuracy and good utilization performance.

The measurement accuracy depends on the following:

* frequency of traffic variation of the incoming traffic, high frequency traffic is more difficult to measure.

* tuning knobs of the measurement mechanism e.g., measuring window size W, sampling time S, and the decision parameters of the estimation approach (updating of the measured value) [BrJa00].

* tuning knobs of the filtering mechanism to smooth the estimated bandwidth.

* processing time of the bandwidth usage measurement, estimation of the measurements can be in a longer or shorter time scale.

* Estimation can be based on the information of the past, this can give us more accurate bandwidth estimation.

* to decrease the estimation errors additional bandwidth MAY be allocated to combat incidental congestion.

* for infrequent traffic, bandwidth measurements are satisfactory, while for frequent traffic, the buffer

occupancy measurements are more efficient."

9. Tunneling

When PHR messages are tunneled within the RMD Diffserv domain, the tunneling messages MUST include the PHR option field.

10. Security considerations

The general security and tunneling considerations stated in Section 6 of [RFC2475] and [RMD-frame] also apply to this PHR.

In addition, unlike Differentiated Services PHBs, the RIMA PHR allows the edge nodes to monitor traffic load status associated with bandwidth or other QoS parameters dynamically. This flexibility makes it more vulnerable to erroneous traffic load of the traffic load status and sabotage. In order to keep functioning properly, the edge nodes MUST be certain that any flow traffic load bandwidth in the network is authorized to do this and only up to that flow's agreed upon limit. If the edge node detects erroneous or malicious behavior, it MUST police that flow to the agreed upon limits or reject it entirely.

Because of the fact that the process of traffic load the traffic load status of a node does not require any reservation state, the RIMA PHR can recover relatively easily from incorrect requests. Thus it is quite safe to deploy the RIMA PHR in a well-controlled network with trustworthy edge nodes.

In order to prevent abuse of the QoS capabilities of the core network, the ingress nodes SHOULD filter any PHR or PDR related header information coming from the outside before sending it through the core network. Whether this information needs to be preserved and later re-inserted or if it should be discarded from the packet or if the entire packet should be discarded is an open issue.

11. References

[BaIt00] Baker, F., Iturralde, C., Le Faucher, F., Davie, B., "Aggregation of RSVP for IPv4 and IPv6 Reservations", Internet draft,

Work in progress.

[BrJa00] Breslau, L., Jamin, S., Schenker, S., "Comments on the performance of measurement based admission control algorithms" Proceedings of INFOCOM 2000, vol 3. p 1233-42, 2000.

[RMD-frame] Karagiannis, G., Rexhepi, V., Westberg, L., Partain, D., Oosthoek, S., Jacobsson, M., Szabo, R., "Resource Management in Diffserv Framework", Internet draft, February 2001 (work in progress).

[RODA] Westberg, L., Karagiannis, G., Partain, D., Oosthoek, S., Jacobsson, M., Rexhepi, V., "Resource Management in Diffserv On DemAnd (RODA) PHR" Internet draft, Work in progress, February 2001.

[RFC791] DARPA INTERNET PROGRAM PROTOCOL SPECIFICATION, "Internet Protocol", IETF RFC 791, September 1981.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC2119, March 1997.

[RFC2205] Braden, R., Zhang, L., Berson, S., Herzog, A., Jamin, S., "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", IETF RFC 2205, 1997.

[RFC2460] Deering, S., Hinden, R., "Internet Protocol, Version 6 (IPv6) Specification", IETF RFC 2460, December 1998.

[RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., Weiss, W., "An Architecture for Differentiated Services", IETF RFC 2475, December 1998.

[RFC2676] Apostolopoulos, G., Willians, D., Kamat, S., Guerin, R., Orda, A., Przygienda, T., "QoS Routing Mechanisms and OSPF Extensions", IETF Experimental RFC 2676, August 1999.

[RFC2859] Fang, W., Seddigh, N., Nandy, B., "A Time Sliding Window Three Colour Marker (TSWTCM)", IETF Experimental RFC 2859, June 2000.

.fi

12. Acknowledgments

Thanks to Pontus Wallentin for reviewing this draft and providing useful input.

13. Authors' Addresses

Lars Westberg
Ericsson Research
Torshamnsgatan 23
SE-164 80 Stockholm
Sweden
EMail: Lars.Westberg@era.ericsson.se

Geert Heijenk
Ericsson EuroLab Netherlands B.V.
Institutenweg 25
P.O.Box 645
7500 AP Enschede
The Netherlands
EMail: Geert.Heijenk@eln.ericsson.se

Georgios Karagiannis
Ericsson EuroLab Netherlands B.V.
Institutenweg 25
P.O.Box 645
7500 AP Enschede
The Netherlands
EMail: Georgios.Karagiannis@eln.ericsson.se

Simon Oosthoek
Ericsson EuroLab Netherlands B.V.
Institutenweg 25
P.O.Box 645
7500 AP Enschede
The Netherlands
EMail: Simon.Oosthoek@eln.ericsson.se

David Partain
Ericsson Radio Systems AB
P.O. Box 1248
SE-581 12 Linkoping
Sweden
EMail: David.Partain@ericsson.com

Vlora Rexhepi
Ericsson EuroLab Netherlands B.V.
Institutenweg 25
P.O.Box 645
7500 AP Enschede
The Netherlands
EMail: Vlora.Rexhepi@eln.ericsson.se

Robert Szabo
Net Lab
Ericsson Hungary Ltd.
Laborc u. 1
H-1037 Budapest
Hungary
EMail: robert.szabo@eth.ericsson.se

Pontus Wallentin
Ericsson Radio Systems AB
P.O. Box 1248
SE-581 12 Linkoping
Sweden
EMail: Pontus.Wallentin@era.ericsson.se

Hamad el Allali
University of Twente
P.O. BOX 217
7500 AE Enschede
The Netherlands
EMail: allali@cs.utwente.nl

Internet Draft Resource Management in Diffserv MBAC PHR October 2002